Informe de Progreso
Técnico y Financiero

Donaciones para la Digitalización de Datos
Red Temática de Especies y Especímenes

Informatización de las colecciones del Museo Argentino de Ciencias Naturales, con énfasis en registros de Parques Nacionales de Argentina

Preparado por: Martín J. Ramírez

12/Jun/2009

Resumen Ejecutivo

Este reporte cubre el periodo 1 de Febrero – 30 Abril, 2009. Se convirtieron registros en papel, o previamente digitalizados, a formato digital de acuerdo a estándares de interoperabilidad. Se reporta un total de 37152 registros (14254 nuevos registros, más 22898 registros previamente digitalizados). Estos registros están integrados a las bases de datos institucionales y son libremente accesibles a través de TapirLink. Los principales desafíos fueron la implementación de flujos de trabajo en paralelo de validación taxonómica y georreferenciación, y afrontar la carga de datos aún cuando ciertos campos no están implementados en la herramienta de captura.

Executive Report

This report covers between 1 February and 30 April, 2009. We converted records in paper, or previously digitized, to a digital format according to standards of interoperability. We report a total of 3715 (14254 new records, pus 22898 records previously digitized). Those records were integrated to the institutional data bases and are freely accessible via TapirLink. The main challenges to overcome were the implementation of parallel workflows for taxonomic validation and georeferencing, and the digitization of data when the appropriate fields are not implemented in the tool to capture data.

1. Resultados de los productos programados y alcances del proyecto

Las bases de datos de las colecciones del Museo se manejan con la aplicación Aurora, desarrollada en nuestra institución. Los campos internos de nuestras bases se mapearon a proveedores TapirLink:

- Colección Nacional de Herpetología (MACN-He, TapiLink: http://168.96.62.13/tapirlink/tapir.php/macnhe). 14254 registros digitalizados. Estos registros fueron tipeados por los data-entry a partir del libro de inventario.
- Colección Nacional de Mastozoología (MACN-Ma, TapirLink http://168.96.62.13/tapirlink/tapir.php/macnma). 22898 registros previamente digitalizados, adecuados a estándares. Estos registros fueron levemente adecuados al formato requerido por la SSTN. En este periodo se produjo la georreferenciación de la mayor parte de los registros.
- Se produjo un *template* MS-Excel para el proceso en paralelo de validación taxonómica, vinculando determinaciones con nombres científicos aceptados.
- Se configuró el servidor de datos para los proveedores TapiLink. Se recurrió a la experiencia de Renato Mazzanti (CENPAT, Trelew, Argentina) en la resolución de problemas de configuración que excedían la documentación del proveedor.

- Se adoptó un protocolo de georreferenciación, levemente modificado de protocolos existentes
- La experiencia adquirida durante este trabajo fue muy útil en las discusiones para delinear el Sistema Nacional de Datos Biológicos de Argentina, de inminente creación[1].

Además de los recursos y avances reportados arriba, se progresó en los siguientes elementos, que formarán parte de los siguientes reportes a medida que cumplan las revisiones internas:

- Conversión de una base de datos de la Colección Nacional de Aracnología, unos 13000 registros que deben adecuarse al formato Aurora y Darwin Core 2. Estado: Avanzado, falta procesar el campo de Notas, en texto libre, que se desglosa en muchos otros campos; pendiente de georreferenciación.
- Georreferenciación de la Colección Nacional de Mastozoología. Estado: 75% completa. Se avanzó en acuerdos de colaboración con un proyecto de la Administración de Parques Nacionales que utilizaría nuestros datos en el corto plazo[2].
- Digitalización de 2400 nuevos registros de la Colección Nacional de Aracnología. Estado: En línea en TapitLink (http://168.96.62.13/tapirlink/tapir.php/macnar), pendiente de validación taxonómica y georreferenciación.
- Digitalización de unos 3000 registros de las Colecciones Nacionales de Ornitología e Invertebrados, que se reportarán en los próximos informes.

2. Metodología empleada y actividades llevadas a cabo para alcanzar los productos programados

Selección de pasantes universitarios. Se realizó un concurso abierto a los estudiantes de la Licenciatura de Biología de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires. Se presentaron 60 postulantes.

Reuniones de seguimiento y seminarios. Todos los lunes se realiza una breve reunión de seguimiento interno seguida de un seminario de actualización abierto al público[3]. En estos seminarios se leen y discuten trabajos científicos metodológicos sobre temas afines a la actividad de los pasantes.

[1] Programa de Organización de Sistemas Nacionales de Grandes Instrumentos y Bases de Datos, http://www.cicyt.mincyt.gov.ar/cicyt_lineas_accion.htm
[2] "Improve the fauna database of the Biodiversity Information Network by adding the data on mammals of special value of the NEA", http://www.oas.org/dsd/IABIN/Component2/Argentina/APN-SM-ValorEspecial.htm
[3] http://www.macn.secyt.gov.ar/Investigacion/proyectos/inv_pro_colecciones_seminarios.php

Digitalización de registros. Los pasantes universitarios digitalizan registros del libro de inventarios o directamente de especímenes de las colecciones, utilizando una interfaz de carga de datos (aplicación Aurora). Cada pasante trabaja 20 horas semanales, distribuidas de acuerdo a su disponibilidad. A los pasantes que digitalizaron directamente especímenes de las colecciones (en vez de libros de inventarios) se les asignaron conjuntos de especímenes preferentemente de Parques Nacionales de Argentina.

Limpieza de registros previamente digitalizados. Las tablas a limpiar se distribuyen en formato MS-Excel con los identificadores únicos originales. Los datos corregidos se tipean en campos de destino conforme a los formatos requeridos. Las modificaciones se incorporan nuevamente a las bases mediante consultas de MS-Access ad-hoc, utilizando los identificadores únicos. Algunas tareas de limpieza de datos son realizadas a nivel administrador directamente sobre las tablas. Dado que es más eficiente procesar bloques enteros del mismo origen, y todavía se está trabajando sobre un bloque grande de más de 13.000 registros, en este primer informe se cubre el cupo comprometido con registros nuevos

Georreferenciación. Se adoptó un protocolo[4] levemente modificado del de MANIS[5]. Los registros de especímenes y localidades fueron obtenidos mediante consultas y preparados en planillas MS-Excel. El georreferenciador agrega los datos de coordenadas geográficas, error, metadatos asociados, y dispone de campos para corregir y uniformar los datos geográficos originales. Los datos se incorporarán a las bases originales del mismo modo que en el ítem anterior.

Validación taxonómica. Los pasantes que ingresan especímenes directamente de la colección de Aracnología tuvieron acceso a la fuente de autoridad taxonómica más actual[6] y coordinaron la consistencia entre nombres y determinaciones con ayuda del curador. Las determinaciones de Mastozoología fueron validadas o actualizadas según la última edición del catálogo de mamíferos[7], en un proceso en paralelo. Se produjeron las planillas para validación taxonómica de Herpetología, y están avanzadas las de datos previos de Aracnología.

Instalación del proveedor TapirLink. Se instalaron Apache Web Server, Intérprete Php, y Mysql server en el servidor de la institución. Se instaló el proveedor de datos Tapirlink, configurando para cada recurso los Metadata, Datasource, Tables, Localfilter, Mapping y Settings requeridos. Se implementaron consultas que mapean los datos de nuestras bases en el formato requerido por el proveedor (DarwinCore).

[4] http://www.macn.secyt.gov.ar/Investigacion/proyectos/inv_pro_colecciones_georef.php
[5] Mammal Networked Information System, http://manisnet.org/
[6] The World Spider Catalog, Version 9.5, http://research.amnh.org/entomology/spiders/catalog/
[7] Wilson & Reader's Mammal Species of the World, 3rd edition, 2005 (http://www.bucknell.edu/MSW3/).

3. Lecciones aprendidas, problemas y soluciones viables

La eficiencia de carga de datos y consistencia de contenidos depende de la existencia de herramientas de carga y manejo de datos. Al inicio de este proyecto en 2008, las herramientas más promisorias (Specify, Ara) estaban en fase de desarrollo, por lo que continuamos desarrollando nuestra propia aplicación (Aurora). La ventaja de una aplicación propia es que tenemos soporte propio, control del desarrollo y la estructura de datos. La desventaja es que hay que implementar cada funcionalidad y mantener el código. Al ser una aplicación pequeña, el aprendizaje es rápido, pero la funcionalidad es limitada.

Problema 1. Campos no implementados. Dado que no existen aplicaciones perfectas, el proceso de digitalización debe realizarse aún cuando ciertos datos no pueden cargarse de la manera ideal. Esto lo hemos encarado implementando un modo estereotipado de colocar los datos en un campo de texto libre, de manera que puedan ser *parseados* eficientemente a medida que los campos son incorporados a la aplicación[8].

Problema 2. Georreferenciación. La aplicación Aurora todavía no soporta todos los datos y metadatos generados por el proceso de georreferenciación (Anexo 1). Mientras se implementan estos campos en la aplicación, estos datos se mantendrán en tablas externas, vinculadas a la base mediante identificadores únicos. Los datos de georreferenciación se incorporarán a la consulta que alimenta al proveedor TapirLink durante Junio.

Problema 3. Validación taxonómica. La aplicación Aurora todavía no soporta un historial de determinaciones para cada registro, ni sinonimias en la tabla taxonómica. Mientras se implementan algunas de estas funciones, los datos dela validación taxonómica (Anexo 2) se tratarán como en el caso de georreferenciación. Los nombres científicos validados se expondrán en en proveedor TapirLink luego de que se implementen las consultas, durante junio de 2009.

La implementación de los esquemas de trabajo sobre tablas externas ha demorado el proceso de georreferenciación y validación taxonómica. En compensación, hemos enfatizado en la digitalización de nuevos registros. Estos valores se equilibrarán a medida que avanza el proyecto.

Notas sobre el tutorial de TapirLink. Para la instalación de datos de TapirLink se siguió el "Manual de instalación y configuración del proveedor", preparado por Ivette Fernández, de Febrero de 2008. La instalación del software requerido por el proveedor no presentó dificultad. Con respecto a la configuración del

[8] "Cómo colocar valores de campos ausentes en Observaciones" y "Mapa Darwin Core - Aurora y abreviaturas para campos no implementados", http://www.macn.secyt.gov.ar/Investigacion/proyectos/inv_pro_colecciones_georef.php

proveedor, sería muy útil contar con un documento donde se muestre un ejemplo completo y funcional, donde pueda observarse el contenido de cada uno de ítems requeridos. En los ejemplos provistos existen campos sin información (solo puede encontrarse la descripción del mismo), y a algunos ítems del ejemplo no puede accederse. Si bien el mapeo de los campos se realizó sin dificultad, al realizar los test, se obtuvieron errores, debido a que los nombres de las bases en MySQL contenían guiones (esto se solucionó eliminando los mismos). En la operación 'search' mediante el uso del Tapirlink XML Client, sería recomendable suministrar un ejemplo y su correspondiente descripción para que el usuario pueda adaptarlo a sus recursos y comprobar su funcionamiento. La adaptación del archivo XML provisto en el cliente es lo que se realizó para verificar el funcionamiento de los recursos agregados del MACN. El tutorial no es explícito acerca del registro en UDDI, podría aclararse que no es necesario para la SSTN.

4. Fondos de Contrapartida (adjunto el Reporte de gastos de contrapartida en tabla de Excel)

Adjuntado como 1raRENDICION MACN-IABIN.xls (rendición original enviada por separado por aminsitradora de fondos)

5. Reporte Financiero (adjunto el Reporte de gastos en tabla de Excel)

Adjuntado como 1raContrapartida MACN-IABIN.xls (rendición original enviada por separado por aminsitradora de fondos)

6. Anexos

Anexo 1. Campos del proceso de georreferenciación
País
Provincia
Departamento
Localidad
Lugar citado
Protocolo
FuenteDeLasCoordenadas
Unidad de las coordenadas
PrecisionDeLasCoordenadas
LatitudDecimal
LongitudDecimal
Datum
DistanciaMaximaDelError
UnidadesDeLaDistancia

ExtensionDeLaEntidad
ObservacionesAcercaDeLasCoordenadas
ObservacionesAcercaDeLaLocalidad
SistemaOriginalDeCoordenadas
NoGoerreferenciadoPorque
GeorreferenciadoPor
FechaGeorreferenciacion
<más varios identificadores internos y valores originales>

Anexo 2. Campos del proceso de validación taxonómica
Phyllum
Clase
Orden
Familia
Genero
Especie
Subespecie
AutorTaxon
CertezaTaxon
FuenteDeValidacion
FechaDeValidacion
ValidadoPor
<más varios identificadores internos y valores originales>